空间流行病学中的疾病制图常用方法

徐 丽[1] 方 亚[2,△]

随着空间分析方法的日益丰富以及局部地理数据可获得性的增加，空间流行病学在对传统流行病学进行拓展的基础上成为系统的流行病学分支[1]。疾病制图(disease mapping)是空间流行病学研究的基本任务，其主要目的在于将疾病危险的空间变异或时空变异在地图上呈现出来[2]，使人们获得直观、感性的认识，为进一步病因学研究或其他研究提供线索。

传统的疾病地图如标点地图、等值区域图(choropleth mapping)[3]等通常基于行政边界，如人口普查和选举病区(electoral wards)在空间上离散地绘制估计的粗率，而研究者普遍认为，疾病相对风险以空间连续的方式度量更为合适[4]；同时，为避免粗率估计不稳定，传统上对于某些没有病例的小区域通常直接进行数据加总，这可能掩盖疾病的真实情况；另外，许多传统地图容易受到各小区域形状与规模不均匀的影响，从而带来视觉偏倚等问题[5]。

近年来，利用空间统计方法绘制疾病地图逐渐成为研究热点之一，其中以地统计和贝叶斯统计为基础的方法占据了研究的主体。这些方法的基本思想是利用“内插”或“平滑”等方法对粗率估计进行处理，以形成易于解释的空间上连续平滑的疾病地图。本文旨在对空间流行病学中的疾病制图常用方法及其应用进行综述，为相关研究提供参考。

“内插”制图法

目前应用较为广泛的“内插”制图法大多基于地统计的基本原理，如距离反比加权、克里格插值、序列指示模拟等。

1. 距离反比加权(inverse distance-weighted，IDW)

IDW 的原则是给予距离近的点的权重大于距离远的点，权重函数是影响绘图结果的关键因素，常用的为距离倒数或距离倒数平方。陆应昶[6]利用 IDW 内插建立江苏省高血压病及其相关区域危险因素的空间分布图，结果发现江苏省 35 岁以上高血压病的分布具有一定的地域性，且与小区域整体的吸烟比率、经济发展水平、受教育程度等变量有一定关联性[6]。

IDW 内插法的优点是简便易行，但其对权重函数的选择十分敏感，且受非均匀分布数据影响大。另外，IDW 假设不同空间位置的病例之间相互独立，且具有相同的概率分布。事实上，不同空间位置的病例之间通常会相互作用，存在着空间相关性。因此，在实际应用中，通常会预先对数据的过离散特征进行处理，然后再利用 IDW 内插法绘制疾病地图，如张志杰[7]利用贝叶斯泊松伽玛混合模型估计中国贵池血吸虫病相对风险(RR)，对过离散形成的虚假衰减变化进行平滑估计，克服了 IDW 容易受非均匀分布数据影响的不足，然后基于获得的贝叶斯 RR 估计值进行 IDW 内插，形成了易于解释的连续平滑的疾病地图。

2. 克里格插值(Kriging interpolation，KI)

KI 也称为空间局部估计或空间局部插值，其最大优点是能够充分利用变量在空间上的自相关特征[8]，是地统计中最为经典的研究方法。该方法建立在变异函数理论基础上，在估计某个待估样本点的数值时不仅考虑落在该样本点的数据，还考虑到邻近样本点的数据以及各邻近样本点与待估样本点的空间相关性与空间异质性，已成为疾病制图的常用方法之一[5,8-12]。

KI 可以理解为广义最小二乘估计，不同的是，它最大限度地利用了样本的空间信息，因此其估计量也满足最佳、线性、无偏的优良性质。它是一种参数估计方法，因此能够对区域计数数据进行解释，如半方差函数的“变程”使得人们能够推断给定时间内某种疾病的传播范围。另外，该方法除了能够给出疾病患病率等变量预测值的平滑曲面，还能够生成预测值的方差图[11]，直观地显示估计的不确定性。但该方法在变异函数估计过程中假定数据同质，这在基于变化的样本容量的情形中不成立。为此，Olaf Berke(2005)[5]提出在应用克里格法之前将经验贝叶斯作为方差稳定变换的方法估计区域风险。另外，该方法还假定空间变量满足二阶平稳性。事实上，疾病变量的空间变异不一定符合这个要求，但 Gotway(2003)[13]已经证明该方法对于非平稳变量的估计效果也很好。

3. 序列指示模拟(sequential indicator simulation，SIS)

Armstrong[14]提出将克里格与蒙特卡罗模拟相结合对空间变量进行插值的方法，即 SIS。SIS 是一种非参数模拟方法，它对变量的分布未做任何假定。理论上，对于每个未被抽样的地点，SIS 估计值合并了邻域

1. 厦门大学经济学院统计系(361005)
2. 厦门大学公共卫生学院，福建省卫生技术评估重点实验室
△通信作者：方亚，Email：fangya@xmu.edu.cn

内可用的所有数据，包括原始数据和所有之前模拟的数值。但在实际应用中，为了简化计算通常仅用邻近的若干个点计算新的模拟值。SIS 的目的是要在研究区域生成许多等概率的实现，从而可以有效地反映异质性造成的不确定性。SIS 被频繁用于描述地下水和土壤中污染物分布的空间格局，描述污染物对人类健康的概率风险[15]。

“平滑”制图法

1. 核估计与等密度投影

核估计(kernel density estimation，KDE)通过从近邻小区域“借力”的方式对变量修匀，从而避免了小区域数据的不稳定性。因此，核估计法能够真实地反映疾病的地理分布，为病因探讨提供重要的线索[16]。

KDE 的主要内容是选择核函数类型与确定最优带宽。常用的核函数类型为均匀核(uniform)、高斯核(Gaussian)、Epanechnikov 核。实际中常用的带宽选择方法为“拇指法则”(rule of thumb)、内插法(plug-in methods)、交叉验证法(cross validation，CV)。其中，Gaussian 与 CV 分别为最常用的核函数与带宽选择方法。通常而言，给定带宽时，核函数类型的不同并不会影响核估计的结果，而带宽的选择则较为关键[8,11]。

通过选择高斯核函数与正态最优化带宽，等密度投影(density equalizing map projections，DEMP)在对基础人口密度异质性调整的基础上对疾病发病率等变量进行平滑，其最终结果是基于地理数据(如人口规模)而不是基于行政边界的地图，从而可以作为协变量进行后续的回归分析。DEMP 能够消除人口密度异质性带来的混杂效应，避免了人为强加的与病因无关的地理边界造成的虚假影响，并且它们提供了风险的连续测量从而避免因小区域病例过少计算出的发病率不稳定问题。另外，基于 DEMP 地图进行的空间分析仅需利用简单灵活的非参数 Kolmogorov 检验，而不用依赖参数的方法评估是否存在空间格局，从而能够更为灵活地分析传染源[17]。有研究者分别依据传统的行政地图与 DEMP 显示旧金山隐孢子虫病例的空间分布，结果发现传统的行政地图表现出明显的病例聚集性，而 DEMP 地图由于考虑到艾滋病与隐孢子虫病的相关性对艾滋病人群分布的异质性进行了调整，从而病例呈现等密度分布[17]。

2. 空间移动平均(spatially moving average，SMA)

SMA 的目的在于对数据进行空间平滑，将其转换成空间上连续的形式，即计算出变量的空间移动平均比率(spatially moving average rate，SMAR)，其在疾病分布的探索性空间分析中很受偏好。空间移动平均法通常采用标化死亡率，如根据普查边界确定的死亡率，来绘制(mapping)健康数据。与标化死亡率不同的是，通过空间平滑，SMAR 不仅去除了个体观测值偏差的影响还消除了特定地点的空间依赖效应，因此，借助该方法绘制的疾病地图对于观察地区的健康状态与提出疾病的病因假说很有用[18]。

SMA 通常与地理信息系统(GIS)结合，可用于消除记录不准确或病例定位错误带来的随机噪音。用于空间流行病学分析的健康与环境数据集通常有多种来源，且不同数据集之间的数据尺度可能不同，因此通常需要对不同数据尺度的数据集进行转换，GIS 提供了解决此类问题的一种途径。但 Mohammad Ali[19] 认为旨在探讨健康与环境的关系、调查疾病空间变异性等研究的方法学过于复杂，阻碍了 GIS 在卫生部门的运用。他认为光栅地理信息系统(raster GIS)是一个可用于空间参照数据简单而实用的工具，能够有效管理和整合多样化数据集，包括卫星图像数据，还可用于创建健康数据的平滑地图。为此，他利用 SMAR 与 raster GIS 对孟加拉国霍乱流行区霍乱发病率[20]与环境危险因素[21]的空间分布特征进行了研究，结果发现当研究环境引起的疾病时，该方法能够降低个体效应的影响，且相较于使用行政边界的方法，创建一个空间平滑的疾病地图更为合理，因为疾病的致病过程通常不与地缘边界相关联。

3. 多项式趋势面模型(trend polynomial surfaces，TPSM)

趋势面模型是指将疾病发病率等变量的空间变异分解为“趋势值”和“剩余值”两部分。其中，“趋势值”用于描述研究区域的系统变异，即可能由环境或人群变化引起的变异，而“剩余值”则用于刻画研究区域内的局部变化。通常采用回归分析的方法拟合趋势面，回归方程的类型很多，但最简单且常用的是多项式回归方程，因此又称为多项式趋势面模型。

趋势面模型阶次的确定是 TPSM 的关键问题。李德云等认为模型阶次的选择取决于趋势面模型检验结果、拟合优度和标准误差的大小等[22]，这也是诸多研究中的常用方法。而薛付忠[23]则认为这是一个复杂的问题，许多方法如直接判定法、拟合优度判定法、剩余均方判定法等都存在一定的优越性和局限性，且它们的结论有时不一致。他认为，应当在遵循地理流行病学原理的前提下，根据疾病空间分布特点，将多种方法综合应用来确定模型的阶次。

TPSM 最初主要用于构建二维曲面，预测变量的空间趋势。薛付忠[24]在二维自回归趋势面模型的基础上加入时间变量，构造三维自回归趋势面模型，不仅可分析肾综合征出血热的空间趋势，而且可预测其空间趋势的时间变化特征。因此，三维趋势面模型是分析预测疾病及其相关因素数据的大范围特征的有用工

具，但其不能用于小范围的细节分析和预测。

TPSM 通常与 GIS 相结合来构造等值线图和二维（或三维）曲面图，最早被用于分析生态数据[25]，现如今已成为疾病空间分析的主要工具之一[22,24,26-31]。另外，王琳娜[32]还尝试将二阶趋势面模型应用于山西省综合医疗服务水平的综合评价。

需要指出的是，由于许多因素的影响，如随机噪声、样本选择偏倚、混杂偏倚等，趋势面分析的结果与疾病空间分布的真实情况往往存在差异，有时甚至会严重歪曲疾病空间分布的真实面目[30]。薛付忠等人研究了边缘效应、调查点不足与共线性等偏倚对趋势面分析结果的影响及相应的控制方法[30,33-34]。

4. 贝叶斯平滑

与核估计类似，贝叶斯方法也能够从邻域“借力”，在保持地理分辨率（resolution）的同时又能够获得稳健估计，成为近年来疾病制图最常用的方法之一。贝叶斯方法在考虑变量的空间自相关性基础上将全局或局部的风险估计作为先验信息，局部估计向全局或邻域的平均水平平滑，由此获得患病率等变量的稳健估计，避免了小群体或小区域极端值的出现[35]。

贝叶斯平滑主要包括经验贝叶斯（EB）与分层贝叶斯（HB）两大类。EB 方法在给定数据情形下首先假定模型参数已知由此获得感兴趣的参数的后验分布，然后据此估计参数。EB 在疾病制图中较为常用[2,5,36-38]，但它存在以下问题：（1）采用迭代估计，收敛速度可能很慢；（2）估计后验方差时没有考虑到由于模型参数估计造成的额外变异，从而无法衡量参数估计的不确定性[39]。

HB 对 EB 进行了改进，其利用后验均值估计参数，后验方差衡量估计的误差，克服了 EB 无法衡量参数估计不确定性的局限。HB 方法容易理解，且其通过分层建模的方式，具有更大的灵活性，在近年来的应用逐渐增多[7,10,39-42]。但该方法通常会涉及高维积分，计算量大。因此，在实际应用中通常采用马尔科夫链蒙特卡罗模拟（MCMC）的方法估计参数，从而避免了计算一个具有高维积分形式的完全联合后验概率公布，而代之以计算每个估计参数的单变量条件概率分布。许多研究者认为，参数先验分布的选择是关键，若使用不恰当的无先验信息分布（如均匀分布），可能导致不合理的后验分布[43-44]，为此，在参数估计之后通常需要进行敏感性分析，以确保估计结果的稳定性。常用的先验分布为高斯分布、均匀分布、伽玛分布等，近年来也有研究者认为，当研究区域或研究的组数较少时，非中心 t 分布或半柯西分布（half-cauchy）可能是更好的选择[45-46]。

事实上，疾病之间可能有共同的病因，因此，通常需要制作疾病的联合地图以初步验证假设是否成立，贝叶斯共同成分模型[47]提供了很好的方法，其结果不仅直观展示了多疾病共同的相对风险，还可以展示每一种疾病特有的相对风险，成为病因探索性分析的有用工具[48]，为多疾病联合干预策略的制定提供指导。

小　结

疾病制图是空间流行病学研究的基本任务，其主要目的在于说明疾病风险的空间或时空变化，为进一步调查提供线索。传统上通常基于估计的粗率进行作图，这在许多情形下不太可靠。因为粗率估计通常是基于小样本，从而标准误差和变异系数都比较大，这在罕见疾病中成为一个尤为明显的问题。因此，通常会使用“内插”或“平滑”技术去除多余的噪声或离群值，从而获得率的稳健估计。

许多“内插和平滑”作图方法，如 IDW、TPSM 等仍然存在着一些缺陷，如基于独立性假设，不能给出估计值的方差，且容易产生过度平滑问题。SMA 法能够消除个体效应与变量的空间自相关性的影响，但无法给出估计值的方差。KI 的最大优点是能够充分利用变量的空间自相关特征且能够给出估计值的方差，但只能获得唯一的“最优”估计。近年来贝叶斯“平滑”方法得到了重大发展，如 HB 法不仅考虑了变量的空间自相关性，还能够给出估计值的方差，而贝叶斯共同成分模型可用于多疾病的联合制图，成为探讨多疾病共同病因的重要方法。需要指出的是，贝叶斯方法也存在“过度平滑”问题，为此，Lawson 和 Clark 建议在贝叶斯动态模型中加入跳跃结构[49]。因此，疾病制图的许多方法各有优劣，在实际应用中通常需要结合使用，如贝叶斯与地统计法相结合用于复杂数据情形下的疾病制图[5,7]。随着流行病学家、医学地理学家等相关研究者对某些疾病地理分布的兴趣越来越大以及计算机技术的进步，空间流行病学中的疾病制图方法将会不断完善。

参 考 文 献

1. Ostfeld RS, Glass GE, Keesing F. Spatial epidemiology: an emerging (or re-emerging) discipline. Trends in Ecology & Evolution, 2005, 20(6): 328-336.
2. Berke O. Choropleth mapping of regional count data of Echinococcus multilocularis among red foxes in Lower Saxony, Germany. Preventive Veterinary Medicine, 2001, 52(2): 119-131.
3. 陈炳为，许碧云，倪宗瓒，等. 地理权重回归模型在甲状腺肿大中的应用. 数理统计与管理，2005，24(3)：113-117.
4. Rushton G. Improving the geographic basis of health surveillance using GIS. GIS and Health, 1998: 63-80.
5. Berke O. Exploratory spatial relative risk mapping. Preventive veterinary medicine, 2005, 71(3): 173-182.
6. 陆应昶，赵金扣，胡晓抒，等. 江苏省高血压病空间地理分布影响因素初探. 中华流行病学杂志，2004，25(7)：91-93.

7. Zhang Z,Carpenter TE,Chen Y,et al. Identifying high-risk regions for schistosomiasis in Guichi, China: A spatial analysis. Acta Tropica, 2008,107(3):217-223.
8. 曹志冬,王劲峰,高一鸽,等. 广州 SARS 流行的空间风险因子与空间相关性特征. 地理学报,2008,63(9):981-993.
9. 王洁贞,薛付忠,马希兰,等. “克立格”定量医学地图的理论方法及其应用. 山东大学学报(医学版),2002,40(2):97-99.
10. Abrial D,Calavas D,Jarrige N,et al. Spatial heterogeneity of the risk of BSE in France following the ban of meat and bone meal in cattle feed. Preventive veterinary medicine,2005,67(1):69-82.
11. Bihrmann K,Nielsen SS,Toft N,et al. Spatial differences in occurrence of paratuberculosis in Danish dairy herds and in control programme participation. Preventive Veterinary Medicine,2012,103:112-119.
12. 康万里,郑素华. 空间扫描统计在中国菌阳结核病分布中的应用. 中国卫生统计,2012,29(4):487-489.
13. Gotway CA,Wolfinger RD. Spatial prediction of counts and rates. Statistics in medicine,2003,22(9):1415-1432.
14. Armstrong MDPA. Theory and Practice of Sequential Simulation. Geostatistical Simulations. Kluwer Academic Publishers, 1993, 111-124.
15. Zeng G,Liang J,Guo S,et al. Spatial analysis of human health risk associated with ingesting manganese in Huangxing Town,Middle China. Chemosphere,2009,77(3):368-375.
16. 王功军,骆福添. 核估计在小地域分析疾病中的应用. 中国医院统计,2005,12(3):231-233.
17. Khalakdina A,Selvin S,Merrill DW,et al. Analysis of the spatial distribution of cryptosporidiosis in AIDS patients in San Francisco using density equalizing map projections(DEMP). International Journal of Hygiene and Environmental Health,2003,206(6):553-561.
18. Ali M,Emch M,Tofail F,et al. Implications of health care provision on acute lower respiratory infection mortality in Bangladeshi children. Social Science & Medicine,2001,52(2):267-277.
19. Ali M,Emch M,Donnay J. Spatial filtering using a raster geographic information system: methods for scaling health and environmental data. Health & Place,2002,8(2):85-92.
20. Ali M,Emch M,Donnay J,et al. The spatial epidemiology of cholera in an endemic area of Bangladesh. Social Science & Medicine,2002,55(6):1015-1024.
21. Ali M,Emch M,Donnay J,et al. Identifying environmental risk factors for endemic cholera: a raster GIS approach. Health & Place,2002,8(3):201-210.
22. 李德云,邓佳云,李津蜀,等. 四川省碘缺乏病趋势面分析模型. 中国地方病学杂志,2004,23(4):58-60.
23. 薛付忠,王洁贞,张际文,等. 疾病空间分布趋势面模型阶次确定方法的研究. 山东大学学报:医学版,2004,42(2):125-130.
24. 薛付忠,王洁贞. 三维自回归趋势面模型在疾病时空动态分析中的应用. 中国卫生统计,1999,16(6):19-22.
25. Gittins R. Trend surface analysis of ecological data. Journal of Ecology, 1968(56):845-869.
26. 王黎霞,刘胜安. 用趋势面分析法研究我国涂阳肺结核患病率的地理分布. 中华流行病学杂志,1995,16(5):274-277.
27. 韩兢,王洁贞,李会庆,等. 山东省主要恶性肿瘤死亡率地域分布的趋势面分析. 山东医科大学学报,2000,38(3):255-257.
28. 罗盛,马峻岭,陈景武. 恶性肿瘤死亡率地域分布的趋势面分析. 中国卫生统计,2008,25(4):357-359.
29. 张际文,王洁贞,薛付忠,等. 山东省糖尿病死亡率的趋势面分析. 山东大学学报(医学版),2003,41(4):388-390.
30. 薛付忠,王洁贞,马吉祥,等. 疾病空间分布趋势面模型的共线性偏倚及其测量与控制. 中国卫生统计,2004,21(2):81-84.
31. 王晓燕,沈毅,陈坤,等. 趋势面分析法在肺癌死亡率地理分布研究中的应用. 中华流行病学杂志,2007,28(6):608-612.
32. 王琳娜,王彤,郭明英,等. 山西省综合医院医疗服务水平趋势面分析. 中国卫生统计,2001,18(1):3-5.
33. 薛付忠,吴晓云,王洁贞,等. 边缘效应偏倚对疾病空间分布趋势面分析结果的影响. 实用医药杂志,2003(10):766-769.
34. 王发银,薛付忠,王洁贞,等. 调查点不足偏倚对疾病空间分布趋势面分析结果的影响. 预防医学文献信息,2004(1):3-6.
35. Clements AC,Pfeiffer DU,Martin V,et al. A Rift Valley fever atlas for Africa. Preventive Veterinary Medicine,2007,82(1):72-82.
36. 黄秋兰,唐咸艳,周红霞,等. 应用空间回归技术从全局和局部两水平上定量探讨影响广西流行性乙型脑炎发病的气象因素. 中华疾病控制杂志,2013(04):282-286.
37. Ettarh R,Galiwango E,Rutebemberwa E,et al. Spatial analysis of determinants of choice of treatment provider for fever in under-five children in Iganga,Uganda. Health & Place,2011,17(1):320-326.
38. 许碧云,陈炳为,李德云. Bayesian 空间泊松模型对小区域非传染病患病率的估计. 中华疾病控制杂志,2010,14(2):166-168.
39. Maiti T. Hierarchical Bayes estimation of mortality rates for disease mapping. Journal of Statistical Planning and Inference,1998,69(2):339-348.
40. Kim DR,Ali M,Thiem VD,et al. Geographic analysis of shigellosis in Vietnam. Health & Place,2008,14(4):755-767.
41. Lee DA comparison of conditional autoregressive models used in Bayesian disease mapping. Spatial and Spatio-temporal Epidemiology,2011,2(2):79-89.
42. Adegboye OA,Kotze D. Disease mapping of Leishmaniasis outbreak in Afghanistan: spatial hierarchical Bayesian analysis. Asian Pacific Journal of Tropical Disease,2012,2(4):253-259.
43. Gustafson P,Hossain S,Macnab YC. Conservative prior distributions for variance parameters in hierarchical models. Canadian Journal of Statistics,2006,34(3):377-390.
44. Gelman A,Jakulin A,Pittau MG,et al. A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics,2008:1360-1383.
45. Gelman A. Prior distributions for variance parameters in hierarchical models(comment on article by Browne And Draper). Bayesian analysis,2006,1(3):515-534.
46. Nathoo F S,Ghosh P. Skew - elliptical spatial random effect modeling for areal data with application to mapping health utilization rates. Statistics in Medicine,2013,32(2):290-306.
47. Knorr Held L,Best NG. A shared component model for detecting joint and selective clustering of two diseases. Journal of the Royal Statistical Society: Series A(Statistics in Society),2001,164(1):73-85.
48. Onicescu G,Hill EG,Lawson AB,et al. Joint disease mapping of cervical and male oropharyngeal cancer incidence in blacks and whites in South Carolina. Spatial and Spatio-temporal Epidemiology, 2010, 1(2):133-141.
49. Toft N,Innocent GT,Mckendrick IJ,et al. Spatial distribution of Escherichia coli O157-positive farms in Scotland. Preventive Veterinary Medicine,2005,71:45-56.

(责任编辑:郭海强)